Identifying Change-points in Biological Sequences via Sequential Importance Sampling

[1]G. Yu. Sofronov and [1]G. E. Evans and [2]J. M. Keith and [1]D. P. Kroese

[1]Department of Mathematics, The University of Queensland, Brisbane, Qld 4072, Australia
E-Mail: georges@maths.uq.edu.au
[2]School of Mathematical Sciences, Queensland University of Technology,
GPO Box 2434 Brisbane, Qld 4001, Australia

Keywords: *Sequential importance sampling, multiple change-point problem, comparative genomics*

EXTENDED ABSTRACT

The genomes of complex organisms, including the human genome, are highly structured. This structure takes the form of segmental patterns of variation in various properties, and may be caused by the division of genomes into regions of distinct function, by the contingent evolutionary processes that gave rise to genomes, or by a combination of both. Whatever the cause, identifying the change-points between segments is potentially important, as a means of discovering the functional components of a genome, understanding the evolutionary processes involved, and fully describing genomic architecture.

One property of genomes that is known to display a segmental pattern of variation is GC content. Genomes are composed of DNA: a long, double-stranded, linear polymer built up from four nucleotide bases, namely adenine, cytosine, guanine, and thymine (A, C, G, and T). The two strands of a DNA molecule form a double-helix, held together by hydrogen bonds formed between G and C nucleotide pairs, and between A and T nucleotide pairs, as illustrated in Figure 1. The two strands are thus complementary; the sequence of either strand can be deduced from that of the other by interconverting G with C, and A with T.

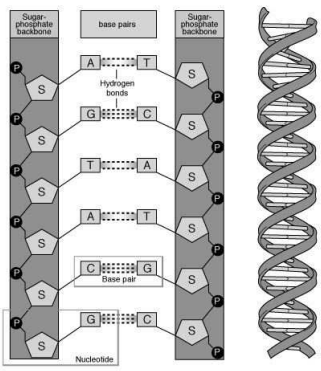

Figure 1. The DNA structure. Source: `http://cnx.org/content/m12382/latest/`

The GC content of a portion of DNA is thus the proportion of GC pairs that it contains. Sharp changes in GC content can be observed in the human and other genomes. For example, Figure 2 shows a small portion of a sequence, in which a sharp increase in GC content is observed at about position 35.

```
1                                   35                                  70
gaattaatatgagatttatatagttgataaagctactccctacccatccccgcctcatctagcccggagg
1000000000101000000000100100000110010111001110011111110100100111111011
```

Figure 2. An example of a sequence and its binary representation with a well defined change-point.

Such change-points may be the boundaries of functional elements, or may play a structural role. We model genome sequences as a multiple change-point process, that is, a process in which sequential data is separated into segments by an unknown number of change-points, with each segment supposed to have been generated by a different process.

Multiple change-point models are important in many biological applications and, particularly, in analysis of biomolecular sequences. For example, multiple change point models can be applied in segmenting protein sequences (which have a 20 character alphabet) according to hydrophobicity. This can aid in the identification of functional domains and can assist in determining the three-dimensional conformations of protein molecules. Another application in which the authors have an interest is in identifying segments that are conserved between two species.

We consider a Sequential Importance Sampling approach to change-point modeling using Monte Carlo simulation to find estimates of change-points as well as parameters of the process on each segment. Numerical experiments illustrate the effectiveness of the approach. We obtain estimates for the locations of change-points in artificially generated sequences and compare the accuracy of these estimates to those obtained via MCMC and a well-known method, IsoFinder. We also provide examples with real data sets to illustrate the usefulness of this method.

1 INTRODUCTION

Eukaryotic genomes display segmental patterns of variation in various properties, including GC content and level of evolutionary conservation. The genome can be divided into segments that are internally fairly uniform with respect to the property of interest, but are significantly different from neighbouring segments. The boundaries of such segments are regions of abrupt change, and are known as *change-points*.

One property that exhibits a segmental pattern of variation is GC content. That is, DNA sequences vary in the local proportion of the nucleotides G and C, as opposed to the nucleotides A and T. Many eukaryote genomes are known to contain structures termed *isochores*, which are large (that is, megabase) segments that differ significantly in GC content from neighbouring segments.

Another important property of genomes that displays segmental variation is the degree of conservation between two species. That is, homologous parts of two genomes may exhibit more sequence similarity than flanking sequences, as a result of selective pressures that inhibit fixation of mutations. One way in which such conserved segments are detected is to first align the two genomes and then perform a "sliding window" analysis — a form of Loess analysis. The alignment is divided into contiguous segments of a fixed width or "window length" and the proportion of matches is determined for each window. Each window thus produces a single data point in the interval [0,1]. The proportion of the genome that is conserved can then be estimated by fitting a mixture model to these data points and estimating the mixture proportion of the most slowly evolving component. This approach has been used to estimate the proportion of the human genome that is conserved relative to mice to be about 5% (Waterson *et al.*, 2002). Such estimates are relevant to current debates about the amount of functional material in genomes and the role of non-protein-coding RNAs.

A problem with using sliding windows is that abrupt changes are blurred over a region equal to the window size. It is possible to reduce the window size, but this increases noise. A more important problem is that this approach is not a very sensitive way to identify statistically significant change-points. A more sensitive technique than sliding windows is sequence segmentation, which involves modelling the sequence as a collection of segments with uniform internal properties and identifying probable locations of change-points. Various methods of genome segmentation have been proposed, and a number of these are reviewed in Braun and Miller (1998). Recent approaches include those of Keith *et al.* (2004) and Oliver *et al.* (1999, 2001, 2002). The authors are actively involved in developing *Markov chain Monte Carlo* (MCMC) approaches to this problem; see Keith (2006) and Keith *et al.* (2007).

In this paper we present a *sequential importance sampling* approach to change-point modelling using Monte Carlo simulation to find estimates of change-points as well as parameters of the process on each segment. We include results of numerical experiments indicating the usefulness of this method. We apply the method to real data from the human genome to detect segmental variation in GC content, but the method could equally be applied to detect segmental variation in other important properties.

The paper is structured as follows. Section 2 includes a statement of the multiple change-point problem in mathematical terms. In Section 3, we explain the basic framework of SIS. In Section 4, we develop SIS for the multiple change-point problem. Section 5 presents the results of two numerical experiments.

2 THE MULTIPLE CHANGE-POINT PROBLEM

Let us formulate the multiple change-point problem in mathematical terms. A binary sequence $b = (b_1, \ldots, b_L)$ of length L is given. A segmentation of the sequence is specified by giving the number of change-points N and the positions of the change-points $c = (c_1, \ldots, c_N)$, where $0 = c_0 < c_1 < \ldots < c_N < c_{N+1} = L$. In this context, a change-point is a boundary between two adjacent segments, and the value c_n is the sequence position of the rightmost character of the segment to the left of the n-th change-point. A maximum number of change-points $N_{\max}$ is specified, where $0 \leq N \leq N_{\max} < L$. The model for the data assumes that within each segment characters are generated by independent Bernoulli trials with probability of success (that is obtaining a "1") θ that depends on the segment. Thus, the joint probability density of $b_1, \ldots, b_L$, conditional on N, $c = (c_1, \ldots, c_N)$, and $\theta = (\theta_0, \ldots, \theta_N)$, is given by

$$\begin{aligned} &f(b_1, \ldots, b_L \mid N, c, \theta) \\ &= \prod_{n=0}^{N} \theta_n^{\mathbb{I}(c_n, c_{n+1})} (1-\theta_n)^{\mathbb{O}(c_n, c_{n+1})}, \end{aligned}$$

where

$$\begin{aligned} \mathbb{I}(c_n, c_{n+1}) &= \sum_{i=c_n+1}^{c_{n+1}} b_i, \\ \mathbb{O}(c_n, c_{n+1}) &= c_{n+1} - c_n - \mathbb{I}(c_n, c_{n+1}). \end{aligned}$$

In other words, $\mathbb{I}(c_n, c_{n+1})$ is the number of ones in the segment bounded by sequence positions c_n+1 and c_{n+1}, and $\mathbb{O}(c_n, c_{n+1})$ is the number of zeros in that same segment.

To formulate the problem in terms of a Bayesian model, a prior distribution must be defined on the set of possible values of $\mathbf{x} = (N, c, \theta)$, denoted

$$\mathcal{X} = \cup_{N=0}^{N_{\max}} \{N\} \times \mathcal{C}_N \times (0,1)^{N+1},$$

with

$$\mathcal{C}_N = \{(c_1, \ldots, c_N) \in \{1, \ldots, L-1\}^N : c_1 < \ldots < c_N\}.$$

We assume a uniform prior both on the number of change-points and on $\mathcal{C}_N$, and uniform priors on $(0,1)$ for each θ_n. Thus, the overall prior $f_0(N, c, \theta)$ is a constant. The posterior density at point $\mathbf{x} = (N, c, \theta)$, having observed $b_1, \ldots, b_L$, is thus given by

$$\pi(\mathbf{x}) \propto \prod_{n=0}^{N} \theta_n^{\mathbb{I}(c_n, c_{n+1})} (1-\theta_n)^{\mathbb{O}(c_n, c_{n+1})}.$$

3 SEQUENTIAL IMPORTANCE SAMPLING

Consider the problem where we wish to evaluate the quantity

$$\ell = \int_{\mathcal{X}} H(\mathbf{x}) \pi(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_\pi[H(\mathbf{X})],$$

where the subscript π means that the expectation is taken with respect to $\pi(\mathbf{x})$ — the target density (in our case the posterior density) — and $H(\mathbf{x}) \geq 0$ is some performance function.

We can then represent ℓ as:

$$\ell = \int H(\mathbf{x}) \frac{\pi(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\, d\mathbf{x} = E_g \left[H(\mathbf{X}) \frac{\pi(\mathbf{X})}{g(\mathbf{X})} \right].$$

We can now get an unbiased estimator for ℓ, called the *importance sampling estimator*, as follows:

$$\hat{\ell} = \frac{1}{N_1} \sum_{i=1}^{N_1} H(\mathbf{X}^{(i)}) \frac{\pi(\mathbf{X}^{(i)})}{g(\mathbf{X}^{(i)})},$$

where $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N_1)}$ is a random sample from a different density g. The ratio of densities

$$W(\mathbf{x}) = \frac{\pi(\mathbf{x})}{g(\mathbf{x})}$$

is called the *importance weight* or *likelihood ratio*.

A variant of the importance sampling technique is know as *sequential importance sampling* (SIS). It is not always easy to come up with an appropriately close importance sampling density $g(\mathbf{x})$ for high-dimensional target distributions. SIS builds up the importance sampling density sequentially.

Suppose that $\mathbf{x}$ can be written in the form $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, where each of the x_i may be multi-dimensional. Then we may construct our importance sampling density as

$$g(\mathbf{x}) = g_1(x_1)\, g_2(x_2 \,|\, x_1) \cdots g_d(x_d \,|\, x_1, \ldots, x_{d-1}),$$

where the g_t are chosen so as to make $g(\mathbf{x})$ as close to the target density, $\pi(\mathbf{x})$, as possible. We can also rewrite the target density sequentially as

$$\pi(\mathbf{x}) = \pi(x_1)\pi(x_2 \,|\, x_1) \cdots \pi(x_d \,|\, \mathbf{x}_{d-1}),$$

where we have abbreviated $(x_1, \cdots, x_t)$ to $\mathbf{x}_t$. The likelihood ratio now becomes

$$w_d = \frac{\pi(x_1)\pi(x_2 \,|\, x_1) \cdots \pi(x_d \,|\, \mathbf{x}_{d-1})}{g_1(x_1)\, g_2(x_2 \,|\, x_1) \cdots g_d(x_d \,|\, \mathbf{x}_{d-1})},$$

which can be evaluated sequentially as

$$w_t = u_t\, w_{t-1}, \quad t = 1, \ldots, d,$$

with initial weight $w_0 = 1$. The *incremental weights* $\{u_t\}$ are given by $u_1 = \pi(x_1)/g_1(x_1)$ and

$$u_t = \frac{\pi(x_t \,|\, \mathbf{x}_{t-1})}{g_t(x_t \,|\, \mathbf{x}_{t-1})} = \frac{\pi(\mathbf{x}_t)}{\pi(\mathbf{x}_{t-1})\, g_t(x_t \,|\, \mathbf{x}_{t-1})}, \quad t = 2, \ldots, d.$$

However this incremental weight requires knowing the marginal probability density functions $\{\pi(\mathbf{x}_t)\}$. This may not be easy and so we need to introduce a sequence of *auxiliary* pdfs $\pi_1, \pi_2, \ldots, \pi_d$ such that (a) $\pi_t(\mathbf{x}_t)$ is a good approximation to $\pi(\mathbf{x}_t)$, (b) they are easy to evaluate, and (c) $\pi_d = \pi$. The SIS method can now be described as follows.

Algorithm 1 (Sequential Importance Sampling)

1. *For each finite $t = 1, \ldots, d$, draw $X_t = x_t$ from $g(x_t \,|\, \mathbf{x}_{t-1})$.*
2. *Compute $w_t = u_t w_{t-1}$, where $w_0 = 1$ and*

$$u_t = \frac{\pi_t(\mathbf{x}_t)}{\pi_{t-1}(\mathbf{x}_{t-1})\, g_t(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})}, \quad t = 1, \ldots, d.$$

3. *Repeat N_1 times and estimate ℓ via*

$$\hat{\ell}_w = \frac{w^{(1)} H(\mathbf{x}^{(1)}) + \cdots + w^{(N_1)} H(\mathbf{x}^{(N_1)})}{w^{(1)} + w^{(2)} + \cdots + w^{(N_1)}},$$

with $w^{(i)} \equiv w_d^{(i)}$ for $i = 1, \ldots, N_1$.

For more on importance sampling and SIS see, for example, Rubinstein and Kroese (2007).

4 SIS FOR THE MULTIPLE CHANGE-POINT PROBLEM

In order to estimate the average GC content, we are interested in the evaluation of the following integral:

$$\ell(y) = \sum_{N=0}^{N_{\max}} \sum_{c \in \mathcal{C}_N} \int_{(0,1)^{N+1}} H(y)\pi(N, c, \theta)\, d\theta,$$
$$y = 1, 2, \ldots, L,$$

where

$$H(y) = \begin{cases} \theta_0, & 1 \le y \le c_1, \\ \theta_n, & c_n < y \le c_{n+1},\ n = 1, \ldots, N-1, \\ \theta_N, & c_N < y \le L. \end{cases}$$

We can represent our change-point variable $\mathbf{x}$ as a $d = N_{\max}$-dimensional vector:

$$\mathbf{x} = (x_1, \ldots, x_d), \quad x_t = (c'_t, \theta'_{0,t}, \theta'_{1,t}),$$
$$1 \le t \le d,$$

where

- c'_t is the position of a change-point, which has been defined at the t-th iteration of Algorithm 2, $(c'_1, \ldots, c'_d) \in \{1, \ldots, L\}^d$;
- $\theta'_{0,t}$ and $\theta'_{1,t}$ are values of the parameter for the segment obtained at t-th iteration to the left and right of c'_t, respectively.

It will also be convenient to define $\theta'_{0,t} = \theta'_{1,t}$, if there is no change-point at the t-th iteration. We rearrange the positions $c'_1, \ldots, c'_d$ in ascending order and denote the resulting positions by $c'_{1,d} \le \ldots \le c'_{d,d}$. If all the positions $c'_{1,d}, \ldots, c'_{d,d}$ are different (in which case all the inequalities are strict) and $c'_{d,d} < L$, then we have d change-points, $c_1, \ldots, c_d$, $0 < c_1 < \ldots < c_d < L$. If there are equalities, then we obtain N, $N < d$, change-points, which we denote by $c_1, \ldots, c_N$, $0 < c_1 < \ldots < c_N < L$. Finally, we take

$$\theta_0 = \theta'_{0,t_1},$$
$$\theta_n = \begin{cases} \theta'_{1,t_n}, & \text{if } t_n > t_{n+1}, \\ \theta'_{0,t_{n+1}}, & \text{if } t_n < t_{n+1}, \end{cases}$$
$$n = 1, \ldots, N-1,$$
$$\theta_N = \theta'_{1,t_N},$$

where $t_k = \min\{t : c'_t = c_k\}$, $k = 1, \ldots, N$, $N \le N_{\max}$.

Next, we build up the proposal density $g(\mathbf{x})$ as follows. Using all observations $I^{(1)} = \{1, \ldots, L\}$, we generate a point $x_1 = (c'_1, \theta'_{0,1}, \theta'_{1,1})$ by simulation from a distribution

$$g_1(x_1) \propto f(b_1, \ldots, b_L \mid x_1) f_1(x_1),$$

where $f_1(x_1)$ is a prior density defined on $\{1, \ldots, L\} \times (0,1)^2$. Recall that we have assumed a uniform prior on $\mathcal{C}_N$ and uniform priors on $(0,1)$ for each θ_n. It follows that $f_1(x_1)$ is a constant.

Then, we have two intervals $I_{0,1} = \{1, \ldots, c'_1\}$ and $I_{1,1} = \{c'_1 + 1, \ldots, L\}$ (one of them may be empty). We choose these intervals proportional to

$$S_{0,1} = \sum_{i \in I_{0,1}} p_2(i, c'_1) \quad \text{and} \quad S_{1,1} = \sum_{i \in I_{1,1}} p_2(c'_1, i),$$

respectively, where $p_2(r_1, r_2)$ is a uniform prior on $\{(r_1, r_2) \in \{1, \ldots, L\}^2 : r_1 \le r_2\}$. Thus, $S_{0,1}$ and $S_{1,1}$ are proportional to the length of the intervals $I_{0,1}$ and $I_{1,1}$, respectively.

Since we have independent observations, $I_{0,1}$ is independent of $I_{1,1}$. So we do not need to know the values of the parameter θ for non-selected intervals. Using the selected interval $I^{(2)} \in \{I_{0,1}, I_{1,1}\}$, we generate the next point $x_2 = (c'_2, \theta'_{0,2}, \theta'_{1,2}) \in I^{(2)} \times (0,1)^2$. Taking into account that the prior distribution on $(0,1)^2$ is uniform, we obtain

$$g_2(x_2 \mid x_1) \propto v^{(2)} \prod_{i \in I^{(2)}} f(b_i \mid x_2),$$

where $v^{(2)} \in \{v_{0,1}, v_{1,1}\}$, $v_{k,1} = S_{k,1}/(S_{0,1} + S_{1,1})$. Similarly, at the t-th iteration we have

$$g_t(x_t \mid \mathbf{x}_{t-1}) \propto v^{(t)} \prod_{i \in I^{(t)}} f(b_i \mid x_t),\ t = 2, \ldots, d,$$

where

$$v^{(t)} \in \{v_{0,t-1}, \ldots, v_{t-1,t-1}\},$$
$$v_{k,t-1} = S_{k,t-1} \left(\sum_{k=0}^{t-1} S_{k,t-1} \right)^{-1}$$

with

$$S_{0,t-1} = \sum_{i \in I_{0,t-1}} p_t(i, c'_{1,t-1}, \ldots, c'_{t-1,t-1}),$$
$$\vdots$$
$$S_{t-1,t-1} = \sum_{i \in I_{t-1,t-1}} p_t(c'_{1,t-1}, \ldots, c'_{t-1,t-1}, i),$$

and

$$c'_{1,t-1} = \min\{c'_1, \ldots, c'_{t-1}\},$$
$$\vdots$$
$$c'_{t-1,t-1} = \max\{c'_1, \ldots, c'_{t-1}\},$$
$$c'_{1,t-1} \le \ldots \le c'_{t-1,t-1},$$
$$I^{(t)} \in \{I_{0,t-1}, \ldots, I_{t-1,t-1}\},$$
$$I_{k,t-1} = \{c'_{k,t-1} + 1, \ldots, c'_{k+1,t-1}\},$$
$$c'_{0,t-1} = 0, \quad c'_{t,t-1} = L,$$

where $p_t(r_1, \ldots, r_t)$ is a uniform prior on $\{(r_1, \ldots, r_t) \in \{1, \ldots, L\}^t : r_1 \leq \ldots \leq r_t\}$.

We may define the sequence of auxiliary pdfs $\pi_1, \ldots, \pi_d$ as

$$\pi_t(\mathbf{x}_t) \propto f(b_1, \ldots, b_L \mid \mathbf{x}_t) f_t(\mathbf{x}_t), \quad t = 1, \ldots, d,$$

where $f_t(\mathbf{x}_t)$ is a uniform prior distribution defined on $\{1, \ldots, L\}^t \times (0,1)^{t+1}$. In particular, $\pi_1(x_1) = g_1(x_1)$ and $\pi_d(\mathbf{x}_d) = \pi(\mathbf{x})$.

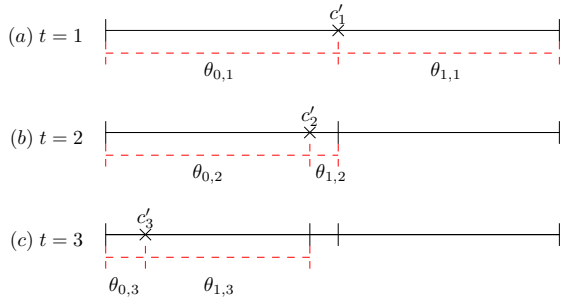

Figure 3. Three iterations of the SIS procedure.

Figure 3 shows the first three iterations of the SIS algorithm for the multiple change-point problem. In part (a) we have a sequence and we draw our first change-point indicated by the cross c'_1. We now draw $\theta_{0,1}$ and $\theta_{1,1}$ for the segments to the left and right of this change-point respectively. In the second iteration (part b) we have picked the left interval proportional to its length as we are using a uniform prior. We draw a change-point c'_2, $\theta_{0,2}$ and $\theta_{1,2}$. This is repeated until we have drawn d change-points.

The SIS for the multiple change-point problem can be defined as the following procedure.

Algorithm 2 (SIS for Multiple Change-point Problem)

1. *Draw $X_t = x_t$ from $g_t(x_t \mid \mathbf{x}_{t-1})$. That is,*
   - *(a) Calculate the weights $v_{k,t-1}$, $k = 0, \ldots, t-1$.*
   - *(b) Select an interval $I^{(t)} \in \{I_{k,t-1}, k = 0, \ldots, t-1\}$ with probabilities proportional to the weights calculated in the previous step. Let $I^{(t)} = \{c_0^{(t)} + 1, \ldots, c_1^{(t)}\}$, $c_0^{(t)}$, $c_1^{(t)}$ are adjacent change-points from $\{c'_1, \ldots, c'_{t-1}\}$ such that $c_0^{(t)} < c_1^{(t)}$.*
   - *(c) Calculate the posterior probabilities*

$$f(c'_t \mid b_i, i \in I^{(t)}) \propto p_1(c'_t) \times \int_0^1 \theta^{\mathbb{I}(c_0^{(t)}, c'_t)} (1-\theta)^{\mathbb{O}(c_0^{(t)}, c'_t)} \, d\theta \times \int_0^1 \theta^{\mathbb{I}(c'_t, c_1^{(t)})} (1-\theta)^{\mathbb{O}(c'_t, c_1^{(t)})} \, d\theta,$$

   where $p_1(c'_t)$ is a uniform prior distribution, $c'_t \in I^{(t)}$.
   - *(d) Insert a new change-point at c'_t (possibly, $c'_t = c_1^{(t)}$) proportional to the probabilities calculated in the previous step.*
   - *(e) Select new Bernoulli parameters $\theta'_{0,t}$ and $\theta'_{1,t}$ for the segments to the left and right of c'_t by sampling from the Beta distribution with parameters $(\alpha_{0,t} = \mathbb{I}(c_0^{(t)}, c'_t) + 1, \beta_{0,t} = \mathbb{O}(c_0^{(t)}, c'_t) + 1)$ and $(\alpha_{1,t} = \mathbb{I}(c'_t, c_1^{(t)}) + 1, \beta_{1,t} = \mathbb{O}(c'_t, c_1^{(t)}) + 1)$, respectively.*

   Let $\mathbf{x}_t = (\mathbf{x}_{t-1}, x_t)$, where $x_t = (c'_t, \theta'_{0,t}, \theta'_{1,t})$.

2. *Compute*

$$u_t = \frac{\pi_t(\mathbf{x}_t)}{\pi_{t-1}(\mathbf{x}_{t-1}) g_t(x_t \mid \mathbf{x}_{t-1})}$$

   and let $w_t = w_{t-1} u_t$, $w_0 = 1$, $t = 1, \ldots, d$.

3. *Repeat N_1 times and estimate ℓ via*

$$\hat{\ell}_w = \frac{w^{(1)} H(\mathbf{x}^{(1)}) + \cdots + w^{(N_1)} H(\mathbf{x}^{(N_1)})}{w^{(1)} + w^{(2)} + \cdots + w^{(N_1)}},$$

   with $w^{(i)} \equiv w_d^{(i)}$ for $i = 1, \ldots, N_1$.

5 RESULTS

In this section we compare our SIS approach to two other methods, IsoFinder (Oliver *et al.*, 2004) and the MCMC approach in Keith *et al.* (2004). For comparison we use two sequences. The first sequence is an artificial sequence with known distribution and the second uses a portion of the *Major Human Histocompatibility region* located on chromosome six.

5.1 Example 1: Artificial data

Let $(b_1, b_2, \ldots, b_{22000})$ be a sequence of independent Bernoulli random variables generated with the parameters given in Table 1. The true profile of this sequence can be seen in Figure 4.

We used the SIS algorithm in Algorithm 2 with $d = 10$, $N_1 = 500$, IsoFinder with a 0.95 significance level and tract size of 1000, and the MCMC algorithm with 100 samples and a step size of 3000. The Mean Squared Error (MSE) is calculated as MSE $= \sqrt{\sum_{i=1}^{22000} (t(i) - e(i))^2}$ where $t(i)$ is the true GC proportion and $e(i)$ is the estimated GC proportion at position i. The results are displayed in Table 2. Figure 4 shows both the MCMC and SIS estimates for the average GC content along the sequence. These two plots are in excellent agreement with each

Table 1. Bernoulli parameters for artificial sequence.

Positions	Bernoulli parameter
1 — 2000	$\theta_0 = 0.35$
2001 — 4000	$\theta_1 = 0.25$
4001 — 6000	$\theta_2 = 0.4$
6001 — 8000	$\theta_3 = 0.5$
8001 — 10000	$\theta_4 = 0.45$
10001 — 12000	$\theta_5 = 0.55$
12001 — 14000	$\theta_6 = 0.4$
14001 — 16000	$\theta_7 = 0.6$
16001 — 18000	$\theta_8 = 0.65$
18001 — 20000	$\theta_9 = 0.5$
20001 — 22000	$\theta_{10} = 0.4$

Table 2. The running time and Mean Squared Error for the three different algorithms when applied to an artificial sequence of 22000 characters.

Algorithm	Time (sec)	MSE
SIS	69	4.36
MCMC	393	2.88
IsoFinder	~ 4	NA

other, supporting the fact that both produced only very small difference between their estimates and the true distribution. It is interesting to note that both algorithms almost always under-estimated or over-estimated the GC content in the same direction. This could be attributed to the fact that although the artificial sequence was drawn from Bernoulli random variables with parameters given in Table 1, it is still a random process and it is possible that the true GC proportions are not exactly the same as the Bernoulli parameters of Table 1.

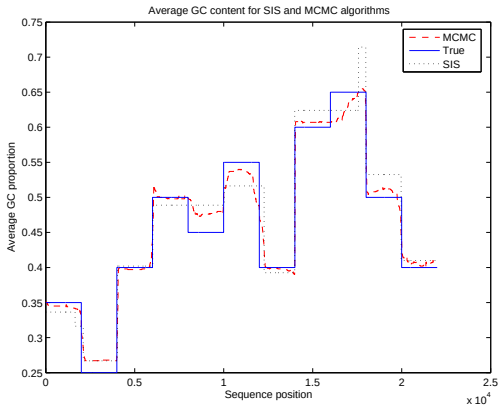

Figure 4. Average GC content as determined by the SIS and MCMC algorithms as well as the true GC profile.

This example illustrates the accuracy of estimates of change-points as well as parameters of the process on each segment.

5.2 Example 2: Real data

The second example uses a real DNA sequence. As a consequence, we do not know the true profile but can still look for agreement between the methods. Using the same algorithm parameters as before, we obtain the results summarized in Table 3.

Table 3. The running time of the algorithms when applied to a segment of the Major Human Histocompatibility region.

Algorithm	Time (sec)
SIS	37
MCMC	582
IsoFinder	~ 4

Here, both the IsoFinder and SIS methods are substantially faster than the MCMC approach. All three methods produced consistent GC estimates and these are shown in Figure 5. To try and gauge the goodness of each method we calculated the sum of the squared difference between each curve. These differences are given in the Table 4.

Table 4. The sum of the squared differences of the profiles for each pair of algorithms.

Algorithms	Difference
SIS - MCMC	4.949
MCMC - IsoFinder	6.1354
IsoFinder - SIS	5.6338

Using these distances as a measure of closeness, the SIS profile is closer to both the MCMC and IsoFinders profiles than the latter two are to each other. This indicates that, on average, the SIS approach produces estimates for the mean GC content that lie in between those of the other methods. When examining Figure 5, it is clear to see that the SIS estimate is similar to both the MCMC and IsoFinder estimates.

6 CONCLUSION

In this paper we have proposed how SIS can be used to identify change-points in biological sequences. The methodology can also be extended to more general multiple change-point models. We have demonstrated the effectiveness of this method in examples using both real and artificial sequences.

For the artificial sequence, our method produced average GC estimates that were in excellent agreement with both the MCMC approach and the true profile. IsoFinder was unable to produce an estimate for this example. The running time of the SIS approach was

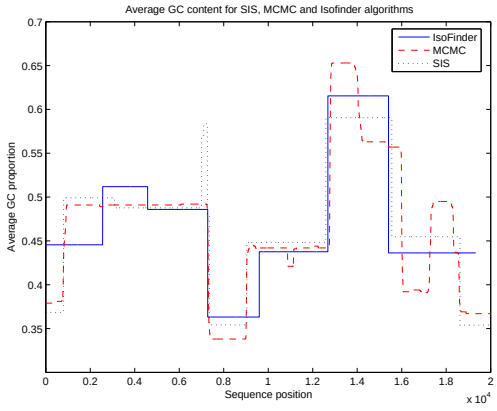

Figure 5. Average GC content as determined by the SIS, MCMC, and IsoFinder algorithms applied to a segment of the Major Human Histocompatibility region.

about five and a half times less than that of the MCMC approach.

When comparing the methods using a segment of the Major Human Histocompatibility region the SIS profile was closer to each of the two established profiles than those were to each other. This indicates that the SIS approach produces estimates for the average GC content that lie in between those of the two other methods. The SIS profile is also similar to both other profiles.

7 ACKNOWLEDGEMENT

G. Yu. Sofronov and D. P. Kroese acknowledge the support of an Australian Research Council discovery grant (DP0556631). J. M. Keith would like to acknowledge the support of Australian Research Council discovery grants (DP0452412, DP0556631) and a National Medical and Health Research Council grant “Statistical methods and algorithms for analysis of high-throughput genetics and genomics platforms” (389892).

8 REFERENCES

Braun, J. V. and Muller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science*, 13, 142–162.

Keith, J., Kroese, D. P. and Bryant D. (2004). A Generalized Markov Sampler. *Methodology and Computing in Applied Probability*, 6(1), 29–53.

Keith, J. M. (2006). Segmenting Eukaryotic Genomes with the Generalized Gibbs Sampler. *Journal of Computational Biology*, 13(7), 1369–1383.

Keith, J. M., Adams, P., Stephen, S. and Mattick, J. S. (2007). Delineating Slowly and Rapidly Evolving Fractions of the Drosophila Genome. To appear in *Bioinformatics*.

Oliver, J. L., Bernaola-Galvan, P., Carpena, P. and Roman-Roldan, R. (2001). Isochore chromosome maps of eukaryotic genomes. *Gene*, 276, 47–56.

Oliver, J. L., Carpena, P., Hackenberg, M. and Bernaola-Galvan, P. `http://bioinfo2.ugr.es/IsoF/isofinder.html`.

Oliver, J. L., Carpena, P., Hackenberg, M. and Bernaola-Galvan, P. (2004). IsoFinder: computational prediction of isochores in genome sequences, *Nucleic Acids Research*, 32 (Web Server issue), W287–W292.

Oliver, J. L., Carpena, P., Roman-Roldan, R., Mata-Balaguer, T. et al. (2002). Isochore chromosome maps of the human genome. *Gene*,300, 117–127.

Oliver, J. L., Roman-Roldan, R., Perez, J. and Bernaola-Galvan, P. (1999). Segment: identifying compositional domains in DNA sequences. *Bioinformatics*, 15, 974–979.

Rubinstein, R. Y. and Kroese, D. P. (2007). *Simulation and the Monte Carlo Method*, 2nd Edition. John Wiley & Sons.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420, 520–562.